January 2020

Decay of URLs Citation: A Case Study of Current Science

Sonia Bansal
soniapta@gmail.com

Seema Parmar

Decay of URLs Citation: A Case Study of Current Science

Sonia Bansal
Assistant Librarian,
Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana
and
Dr. Seema Parmar
Assistant Librarian,
Nehru Library, CCS Haryana Agricultural University, Hisar
E-mail: seemaparmar9@gmail.com

Abstract

The present study is conducted to analyze the accessibility, deterioration and half-life of URLs of web documents cited in Current Science Journal published during 2015 - 2016. A total of 1724 URLs cited in the 1564 articles were examined. It was found that 56.67 percent of URLs were accessible at the time of testing and the remaining 43.33 percent of URLs were not accessible. Out of all HTTP error messages, HTTP 404 – 'file not found' was the irresistible error message encountered and represented 59.03 percent of all HTTP error messages. Average half-life of URLs of missing URLs was estimated to be 1.76 years. Even though there are various retrieval tools being used to recover vanished URLs, still there is a need to improve such tools.

Keywords: URLs, citations, cited articles, corrosion, e-resources, e-publishing, e- journals.

1. INTRODUCTION

In the era of ICT when information is exploding in every aspect of life, academic sector is no exception. There is explosion of information in terms of various types of documents available on web in different formats. Nowadays, when e-content is being preferred for reading by many research scholars, e- publishing of journals, books and other documents is increasing day by day. Likewise, the number of citations from web resources has also increased under the influence of e-resources. 'Citation has several important purposes: to avoid plagiarism, to attribute prior or unoriginal work and ideas to the correct sources, to allow the reader to determine independently whether the referenced material supports the author's argument in the claimed way, and to help the reader gauge the strength and validity of the material the author has used (Gul, et al, 2014). The citing behavior of researchers has been influenced by the Web and this has subsequently influenced the growth of web citations (Moghaddam et al, 2010). Citations from not only 'research articles' but from other documents are also increasing like e-ppts, e-lectures, videos, and social media tools etc. The use of citations or URLs of other type of e-resources instead of e-journals have several issues of access after a period of time like personal homepages are likely to disappear, location of researcher is changed, reconstruction of some websites without maintaining the old links, changing the protocol portion of the URL from FTP servers to HTTP. Despite the popularity of web citations, it is clearly reflected from some of recent studies that URL

corrosion problem is a very common issue, not for webmasters only, but also for scholar community. This study makes an attempt to analyse the accessibility and decay of web citations cited in the articles of Current Science journal.

1.1 OBJECTIVES

The study has been conducted keeping in view the following objectives:

- To identify the total output and citations used in articles of Current Science journal during 2015-16
- To find out year-wise distribution of web and print citations.
- To ascertain active/missing web citations.
- To find out error codes associated with missing web citations.
- To identify domains associated with web citations.

2. RESEARCH METHODOLOGY

The study has been conducted to know the accessibility and decay of web citations reported in the published articles of Current Science journal during the period 2015-16. For collection of data, every issue of the journal was consulted personally. A total of 1724 web citations were identified from 1564 articles. All these web citations/URLs were checked from W3C link checker (http://validator.w3.org/checklink) to check their accessibility. Book reviews and editorials were not considered for the purpose of this study.

3. REVIEW OF RELATED LITERATURE

Prithviraj and Kumar (2014) studied accessibility, corrosion and half-life of URLs cited in the articles of Indian LIS conference proceedings published from 2001 to 2010 and found 5,698 URLs cited in 1700 articles. About 50.09 percent of URLs were not found working and 49.91 percent of URLs were accessible during testing period, average half-life of URLs of missing URLs was estimated to be around five years. Gul, Mahajan and Ali (2014) analyzed the growth and decay rate of URL citations cited in one of the eminent information web magazine ARIADNE during 2010- 2012 and found that majority of errors were due to the missing content (http 404-file not found) representing 52.68% of all http error codes, followed by “http 500” (24.73%) and “http 403” (19.35%). The domain “.com/.co” was most stable and persistent domain with 95 per cent accessibility. The greatest number of web

resources cited in the articles, were found to be of “html” and “htmls” formats and “ppt” files were found to be most stable with 100% accessibility. P. Habibzadeh (2013) investigated the trend of citation to URLs in five general medical journals from 2006 to 2013 to compare the trends in mainstream journals with small journals. He studied a total of 2822 articles. Since 2006 onwards, the number of citations to URLs increased in the journals (doubling time ranged from 4.2 years in EMHJ to 13.9 years in AIM). Overall, the percentage of articles citing at least one URL has increased from 24% in 2006 to 48.5% in 2013. Accessibility to URLs decayed as the references got old (half life ranged from 2.2 years in EMHJ to 5.3 years in BMJ). The ratio of citation to URLs in the studied mainstream journals, as well as the ratio of URLs accessible were significantly ($p<0.001$) higher than the small medical journals. Spinellis (2003) examined the accessibility and decay rate of references by extracting and inspecting 4224 ULR references from 2471 computer science articles published during 1995-99. He found 27% URLs not accessible and 50% of them became inaccessible 4 years from the date they were published. Sife and Bernard (2013) while examining the persistence and decay of web citations in theses and dissertations of Sokoine National Agricultural Library found that 58% web citations were inaccessible, error message ‘404 File Not Found’ was encountered to 92.7%, domain ‘.com’ contained 28.2% missing URL and average half-life for the URLs was 2.5 years.

4. DATA ANALYSIS

Data extracted from Current Science journal for present study was put into following tables for data analysis and interpretation:

4.1 Distribution of web and print citations in Current Science

It is very apparent from the table 4.1 that a total of 33954 citations were reported in the 1514 articles published in Current Science journal during 2015 and 2016. Out of total citations, only 5.07 were web citations and 94.92 were print citations. The number of articles published in the year 2015 was more than the year 2016. The average citations per article were 21.96 in the year 2015 and 21.45 in 2016. The average web citations per paper in 2015 were 1.14 and in 2016 was 1.13, while print average citation per paper in 2015 was 21.99 and in 2016 was 20.56.

Table.1

Year-wise distribution of web and print citations in Current Science

Year	No. of Articles	Total Citations	Average Citations per Article	Web Citations	Average Web Citation per Article	%age	Print Citations	Average Print Citation per Article	%age
2015	804	17654	21.96	870	1.08	4.92	16784	20.87	95.07
2016	760	16300	21.45	854	1.12	5.23	15446	20.33	94.76
Total	1564	33954	21.71	1724	1.10	5.07	32230	20.61	94.92

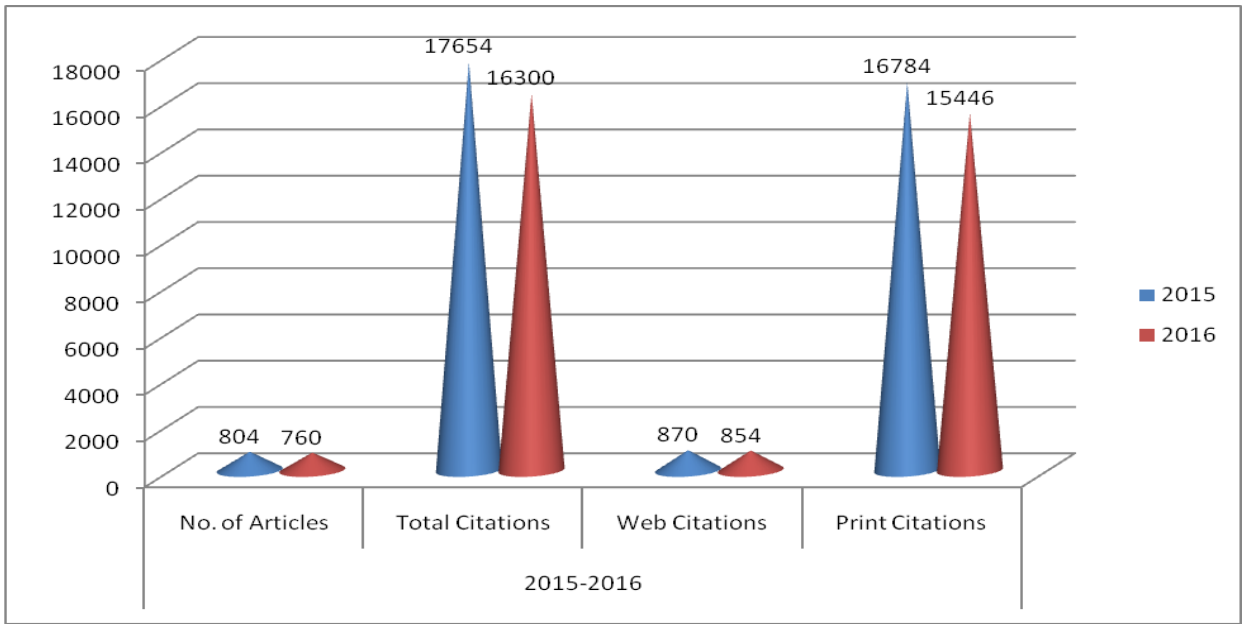

Fig.1

4.2 Status of Active and Missing Web Citations

Table 4.2 depicts the depicts the status of active and missing web citations out of total web citations mentioned with total articles during 2015-16. It is very clear that out of total web citations a little more than half of the citations (56.67%) were active, while rest were found inactive or not working (43.33%). The articles published in 2015 were having only 49.65 percent active web citations while publications of 2016 were having more active web citations (63.82%). It is very clear that missing or inactive web citations were more in 2015 than 2016.

Table.2

Active and missing web citations

Year	Total Web Citations	Active Citations	Percentage of Active Citations	Missing Citations	Percentage of Missing Citations
2015	870	432	49.65	438	50.35
2016	854	545	63.82	309	36.18
Total	1724	977	56.67	747	43.33

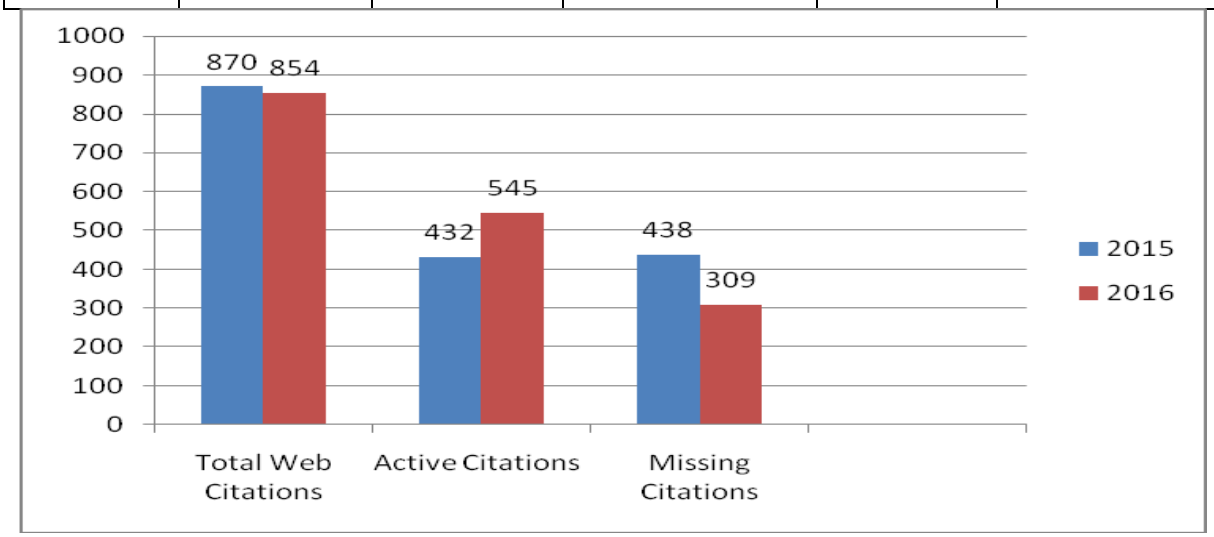

Fig.2

4.3 HTTP errors associated with missing web citations

Table 4.3 represents the distribution of HTTP error codes associated with missing web URLs. Out of total inactive URLs, error type 404 accounted for more than half of missing citations' error (59.03%), followed by other HTTP errors, HTTP 500 (29.72%), HTTP 403 (8.30%). Rest six HTTP errors collectively accounted for 2.95% of missing URLs.

Table.3

HTTP errors associated with missing web citations

Year	HTTP 302	HTTP 400	HTTP 403	HTTP 404	HTTP 410	HTTP 500	HTTP 503	HTTP 504	HTTP 523	Total
2015	1	9	31	245	1	150	1	0	0	438
2016	1	3	31	196	0	72	2	2	2	309
Total	2	12	62	441	1	222	3	2	2	747
%age	0.27	1.61	8.30	59.03	0.13	29.72	0.40	0.27	0.27	

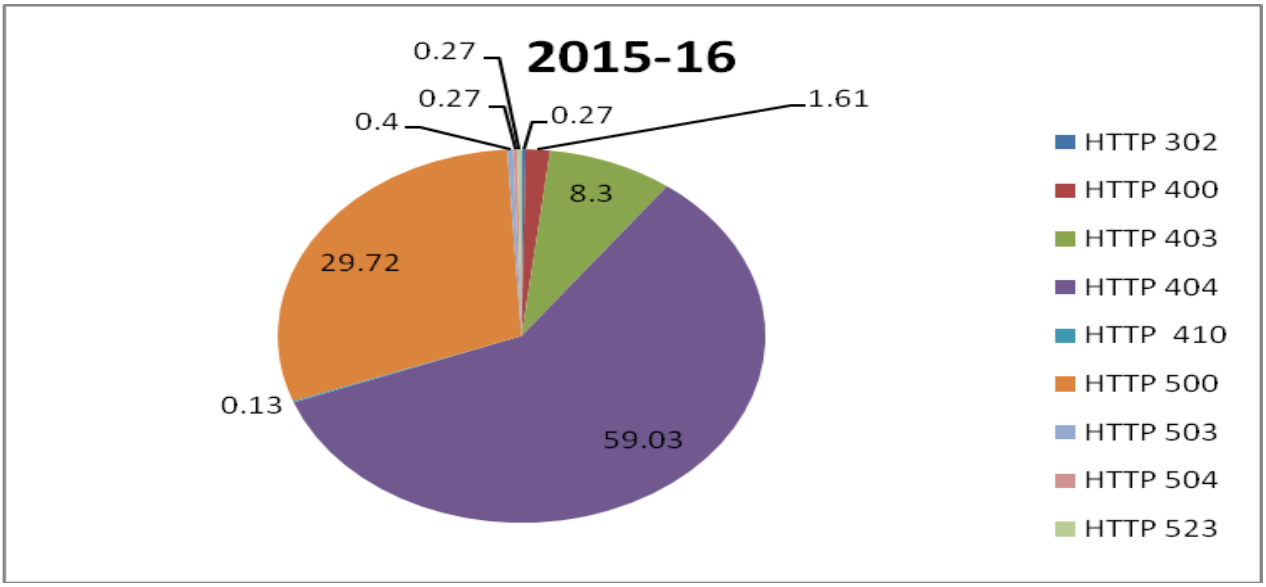

Fig.3

4.4 Web Citations by Domain Type

In this study, only six domains viz. org, .gov, .com, .edu, .ac and .net were taken into consideration; however, those domains not belonging to any of these categories were included in "others". Table 4.4 reveals that 35.21% URLs were associated with .org domain, followed by .com (19.90%), .gov (13.05%). It is also very apparent that the web citations which were associated with domain .net were containing more active links, followed by domain .com (68.80%)

Table.4

Distribution of web citations by domain type

Domains	Total No. of URLs in 2015	Total No. of URLs in 2016	Total No. of URLs	%age	No. of Active URLs	%age of Active URLs	No. of Missing URLs	%age of Missing URLs
.org	331	276	607	35.21	384	63.26	223	36.74
.gov	104	121	225	13.05	84	37.33	141	62.67
.com	162	181	343	19.90	236	68.80	107	31.20
.edu	45	39	84	4.87	50	59.52	34	40.48
.ac	23	28	51	2.96	28	54.90	23	45.10
.net	13	26	39	2.26	29	74.36	10	25.64
Others	192	183	375	21.75	166	44.27	209	55.73

Total	870	854	1724	100.00	977	56.67	747	43.33

4.5 Half-life of Web Citations

The half-life is the time required for exactly half of the web citations in articles to decay. The procedure adopted by Koehler (1999); Tyler and McNeil (2003); Dimitrova and Bugeja (2007); Mardani and Sangari (2013) has been used to calculate half-life of URLs.The half-life of URLs (t_h) has been calculated using following formula:

$$t_h=[t \ln(0.5)] / [\ln W(t) – \ln W(o)]$$

where t_h is the estimated number of years it takes for 50 percent of the web citations to stop working, W(o) is the number of working online citations at the time of publication, W(t) is the number of working online citations at some later time t. Using this formula, half-life has been calculated and the data is presented in Table 4.5. The average half-life for the missing URLs was estimated to be 1.98 in the year 2015 and 1.54 years in the year 2016. For both the years average half-life for the missing URLs was estimated to be 1.76. This means that it will take about 1.76 (approximately 2 years) for half of the URL citations to vanish.

Table 4.5

Half-life of Web Citations

Year	Time (t)	Total no. of URLs W(o)	No. of Active URLs W(t)	Half-life
2015	2	870	432	1.980229
2016	1	854	545	1.543258
Both years		1724	977	1.7617435

5. SUMMARY AND CONCLUSION

The study reflected that many web resources cited in the current Science Journal have disappeared from their original locations. The most common error message was 404 File Not Found and the .org top-level domain had the highest number of missing URLs. It is also found that average half-life for the missing URLs is 1.76 which means that it will take approximately 2 years for half of the URL citations to vanish. This lack of constancy of web

references implies that ever availability of online information resources is not sure or guaranteed. There are some solutions which can reduce the rate of decay of URLs- i.e Authors' responsibility to not commit typing error while typing URLs, checking URLs before including in the article, maintaining digital copies of cited web resources. Authors can mention DOIs in reference list along with URLs, e-journals and databases of some specific domains (rate of corrosion of URLs of which is comparatively less) may be considered for citations. Institutional Repositories may also be used to cite publications.

6. REFERENCES

- Gul, S., Mahajan, I., & Asifa A. (2014). The growth and decay of urls citations: A case of an online library & information science journal. Malaysian Journal of Library & Information Science, 19 (3), 27-39
- Moghaddam, A.S., Saberi, M.K., & Esmaeel, S.M. 2010. Availability and half-life of web references cited in information research journals: A citation study. *International Journal of Information Science and Management*, 8(2), 57-75.
- Prithviraj, K., & Kumar, B. S. (2014). Corrosion of URLs. *IFLA Journal*, 40(1), 35-47.
- Habibzadeh, P. (2013). Decay of references to web sites in articles published in general medical journals: Mainstream vs small journals. *Applied Clinical Informatics*, 4(4), 455–464.
- Spinellis, D. (2003). The decay and failures of web references. *Communications of the ACM*, 46(1), 71–77.
- Sife, Alfred S., & Bernard, R. (2013). Persistence and decay of web citations used in theses and dissertations available at the Sokoine National Agricultural Library, Tanzania. *International Journal of Education and Development using Information and Communication echnology*, 9(2), 85-94.
- Koehler, W. (1999.) An analysis of web page and web site, constancy and permanence. *Journal of the American Society for Information Science*, 50(2), 162–180.
- Tyler, D., & McNeil, B. (2003). Librarians and link rot: a comparative analysis with some methodological considerations. *Portal: Libraries and the Academy*, 3(4), 615–632.

- Dimitrova, D.V., & Bugeja, M. (2007). The half-life of Internet references cited in communication journals. *New Media and Society*, 9(9), 811–826.
- Mardani, A., & Sangari, M. (2013). An analysis of the availability and persistence of web citations in Iranian LIS journals. *International Journal of Information Science and Management*, 3(1), 29–42.